移动互联网用户终端换机预测的研究与实现 *

符　静，张治中，陈粤龙

(重庆邮电大学 通信网与测试技术重点实验室，重庆 400065)

摘　要：为解决预测潜在换机用户的低效率与实际应用问题，设计并搭建基于大数据平台的换机预测系统。该系统首先采集通信网络各接口的数据并收集外部数据；再通过解析处理平台对网络接口数据进行分发、解码、合成、关联等处理，对外部数据进行 ETL 处理，然后将处理后的数据存入 HDFS 中；进一步，在大数据平台上应用 Spark 组件建立基于逻辑回归的换机预测模型，输出潜在换机用户；最后，选取了某西部城市部分用户数据进行系统测试，所得结果表明，该换机预测系统的预测准确率为 71%，可以较好地识别出潜在换机用户，为运营商及手机制造商的精准营销提供可靠支撑。

关键词：移动互联网；换机预测；逻辑回归；大数据

中图分类号：TN929.5　　doi: 10.3969/j.issn.1001-3695.2017.12.0781

Research and implementation of users terminal replacement prediction in mobile internet

Fu Jing, Zhang Zhizhong, Chen Yuelong

(*Communication Networks Testing Engineering Research Center*, *Chongqing University of Posts & Telecommunications*, *Chongqing* 400065, *China*)

Abstract: In order to solve the low efficiency and practical application of predicting the potential phone replacement user, this paper designed and built a phone replacement prediction system based on big data platform. This system firstly captured signaling data from multiple network interface and collected external data. Through the parse platform, the data from network interface would be distributed, decoded, synthesized and correlated, and the external data would be processed by ETL tools, and then storing processed data into HDFS. Further, the paper established a phone replacement prediction model, which based on logistic regression, using spark components in the big data platform and output the potential phone replacement users. Finally, the paper chose part of the western city's user data for system testing. The result shows that the prediction accuracy of the phone replacement prediction system is up to 71%. It can preferably recognize potential phone replacement users, and provide reliable support for the precise marketing of operators and mobile phone manufacturers.

Key Words: mobile internet; replacement prediction; logistic regression; big data

0　引言

移动互联网用户在使用手机的过程中产生大量的信息，运营商积累了这些用户的相关信息，如手机机龄、流量使用情况、异常开关机次数等，但在海量信息中，运营商和用户都很难及时交互双方所需要的信息。对于运营商而言，无法知道哪些用户有潜在的换机需求、需求的手机类型及可接受的价位等信息，从而做针对性的销售，即精准化营销[1-2]。对于用户而言，虽然市场上手机种类繁多，但是不知道有哪些手机更适合自己，而且性价比更高。运营商与手机用户之间缺乏有效沟通，因此运营商有必要对用户业务和流量等信息进行全面系统的研究分析，以此挖掘出潜在换机用户，这不仅有利于运营商扩大用户市场，增加经济效益，还有利于用户获得更好的体验。

用户终端换机预测是针对运营商数据以及用户上网数据、网络接口采集数据等各类数据，通过数据挖掘方法进行分析，寻找数据间的隐藏关系，从大量终端用户中识别出有换机趋势的用户，便于向潜在换机用户进行精准营销。现有的换机预测研究较少，刘力凯，王国胤，邓维斌[3]利用基于优势关系粗糙集方法对有换机意向的用户进行分类选择；刘畅[4]基于 Cox 回归模型研究发现影响终端换机的因素，其结果表明影响用户换机行为的因素包括年龄、性别、终端品牌等；熊冰妍，王国胤，邓维斌[5]对决策树算法进行改进，并提出了一种基于分级式决

基金项目：国家科技重大专项资助项目（2015ZX03001013）；教育部-中移动科研基金资助项目（MCM20150508）；重庆市重点产业共性关键技术创新重大主题专项资助项目（cstc2017zdcy-zdzx0030）；重庆高校创新团队资助项目（KJTD201312）

作者简介：符静（1992-），女，重庆人，硕士研究生，主要研究方向为通信网测试技术、数据挖掘分析（fj_anne@foxmail.com）；张治中（1972-），男，教授，博士，主要研究方向为第四代移动通信测试技术、宽带信息网络；陈粤龙（1994-），男，硕士研究生，主要研究方向为通信网测试技术、数据分析.

策树的换机预测方法；Yang 等人[6]从手机用户使用 APP 的数据入手，应用生存分析模型预测用户是否换机。

现有换机预测研究多是从理论算法上进行的，且其模型分析的数据仅来源于运营商；而本文提出了基于 Hadoop 平台的换机预测系统，在 Spark 组件上应用逻辑回归算法建立换机预测模型,挖掘潜在换机用户。此外，换机预测模型的数据源不仅有运营商的数据，还包括从 2/3/4G 各网络接口采集的数据。

1　换机预测系统设计

为了能够更好的解决用户需求并拓展运营商业务，本文设计并搭建了基于 Hadoop 平台的换机预测系统。换机预测系统包括数据采集、数据处理、数据挖掘、应用展示等功能模块。基于 Hadoop 平台的换机预测系统框架如图 1 所示。

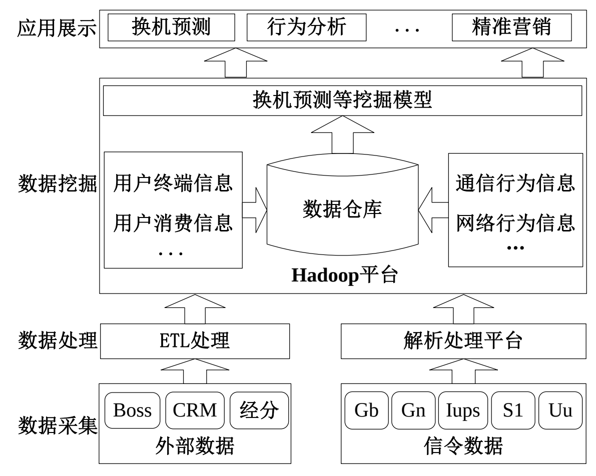


图 1　换机预测系统框架

1.1　数据采集处理

换机预测的原始数据通过数据采集层获取，主要包括外部数据与信令数据两部分。外部数据包括 Boss 系统、CRM 系统以及经分系统的数据，主要通过 ETL 进行处理，将外部数据加载到 Hadoop 数据仓库中去。信令数据由采集卡等设备实时采集获得，包括 Gb、Gn、Iups、S1、Uu 等网络接口的数据，再通过解析处理平台对其进行处理，处理后所得数据传送到 Hadoop 平台。解析处理平台包括数据分发解码、合成 CDR、DPI 识别、关联出表四个模块，其流程如图 2 所示。

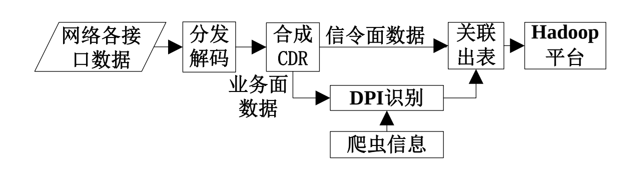


图 2　解析处理平台

1.1.1 数据分发解码

该模块主要完成对所采集网络接口数据的解码功能。首先识别数据类型，判断数据是信令面还是业务面的数据；再根据数据类型的不同分发到相应的解码系统；最后通过解码系统对网络接口数据进行解码，包括关键字段的提取与详细比特内容的“翻译”。

解码系统根据接口协议栈进行解码，协议栈解码都是从底层依次向上层进行的，识别协议栈的每层协议后，调用相应的解码函数进行解码。下面以 S1-U 接口数据的解码为例，说明数据解码流程。首先，对于 S1-U 接口数据，判断数据是否为空，如果不为空，才进行解码；再识别协议栈的最底层协议（IP 协议），调用相应解码函数（IP_decode()）进行解码，IP 解码函数如下。

```
int Ipv4_FDecode(IN const void * pData, IN const int32 nBitLen, IN void * pContext, IN void * pDetail, OUT SDUINFO ** ppSduInfo)
{
uint8 * pDataHead = (uint8 *)pData;
int32 nLength = nBitLen;
......}
```

在底层数据（IP 协议数据）解码完成后判断是否存在上层数据(UDP 协议数据），如果存在，再根据上层协议调用相应的解码函数（UDP_decode()）进行解码，依次循环直至解完原始数据；再调用 FillTreeBuf 函数，构建树型结构，填充解码结果。

1.1.2 合成 CDR

CDR 合成模块主要是对数据解码后的结果进行合成，并在内存中记录 CDR 和原始消息的对应关系，其具体流程如图 3 所示。

首先从解码结果消息中提取共有关键信息 Key；其次，判断 Key 对应的 CDR 记录是否存在，如果存在，就从 hash 表中获取 key 对应的 CDR，并更新 CDR 属性信息；如果不存在，就建立 hash 索引和新的 CDR 记录，并设置 CDR 属性信息。然后，判断当前消息是否为 CDR 结束消息，如果不是，后续消息将完善 CDR 内容；如果是，则移除 Key 并把完整的 CDR 传送到出表模块。

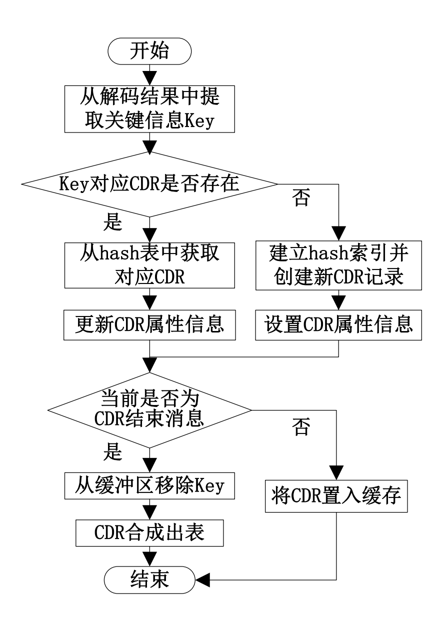


图 3　CDR 合成处理流程

1.1.3 DPI 识别与关联出表

DPI 识别模块主要是识别用户访问互联网的数据，包括业务类型，如视频、购物、阅读等具体内容，此模块需要爬虫信息库的支持，爬虫信息库中存储爬虫程序爬取的一些信息，主要是各网站的业务关键信息，如视频名称、主演等信息。

该模块的输入是 CDR 合成文件数据，首先需要加载爬虫信息库，并读取 CDR 合成文件，再根据 IP 五元组信息，即源 IP 地址，源端口，目的 IP 地址，目的端口和协议类型来查找业务类型；依次通过爬虫信息库、Host、URL 与其他特征来识别业务，成功则填充识别统计表，失败则结束，关键代码如下。

```
switch(m_arr Level[i])
{
case DPI_LEVEL_IP:
{
if(false == m_Check Stat[DPI_LEVEL_IP].Is Ok())
{ Search Ch Feature(dpi Info,Result Info ,context ,DPIText,FEATURE_LEVEL5_IP);
} } ......
default: break;
} }
```

关联出表主要是为了补充并完善 CDR 记录，对合成后的业务面 CDR 记录中的空缺信息，通过信令面的相互关联信息进行交叉回填，回填后将完整的 CDR 合成结果以 CSV 文件格式存盘，实时保存 CDR 的合成记录，便于上层使用。

1.2 数据挖掘与应用展示

数据挖掘模块是在 Hadoop 平台上完成数据导入、统计、分析、预测等功能，其核心是建立相应的数据挖掘模型。应用展示模块的主要功能是对大数据平台处理的数据结果进行展示，根据专题的不同，分为用户换机预测、用户行为分析和用户套餐推荐等。

数据挖掘包括数据准备、数据预处理与数据建模三个阶段，其具体设计如图 4 所示。

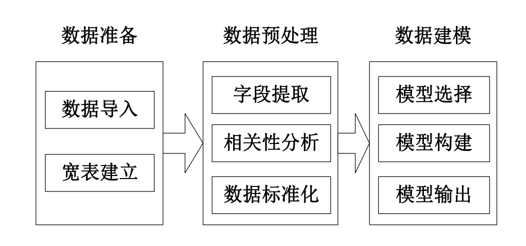


图 4 数据挖掘过程

数据准备阶段包括数据导入、宽表建立等工作。对于解析处理平台与 ETL 处理所得到的各类数据，导入到 Hadoop 平台的 HDFS 中，HDFS 是分布式文件系统[7]，专用于存储各类文件。导入的数据是涵盖通信行为信息、网络行为信息、用户基本属性和用户消费信息等不同维度的数据。然后建立 boss、信令等数据基础宽表。

数据预处理是对基础宽表数据进行处理，根据专家经验法提取特征字段，并对字段进行两两相关性分析，去除相关性较大的字段，形成模型的输入字段；进一步，对数据进行数据质量检查、变量转换等标准化处理，得到输入模型的数据宽表。

数据建模主要分为模型选择、模型建立与模型输出三阶段。第一阶段，根据业务目标选择合适的数据挖掘模型，业务目标是实现换机预测系统，因此换机预测模型选择为逻辑回归预测模型。第二阶段，通过数据挖掘算法进行建模，本文基于 spark 组件建立逻辑回归模型，对数据进行换机预测挖掘。第三阶段，输出换机预测模型结果，即输出潜在换机用户。

2 换机预测模型

换机预测模型主要是采用逻辑回归（Logistic Regression）算法[8-11]在 spark 组件上建立模型。逻辑回归是通过 N 个影响因素预测变量发生的概率，多用于二分类情况。本文需要预测的是客户是否为潜在换机用户，是典型的二分类变量（1 表示是，0 表示否），因此基于逻辑回归算法建立换机预测模型。

逻辑回归算法

逻辑回归的思想也是基于线性回归，属于广义线性回归模型[12-13]。线性回归的公式如(1)所示。

$$z=\beta_0+\beta_1 x_1+\cdots+\beta_n x_n=\boldsymbol{\beta}^{\mathrm{T}} x \tag{1}$$

sigmoid 函数公式如式(2)所示。

$$\sigma(x)=\frac{1}{1+e^{-x}} \tag{2}$$

逻辑回归是将线性函数的结果映射到了 sigmoid 函数中，如式(3)所示。

$$h_\beta(x)=\frac{1}{1+e^{-z}}=\frac{1}{1+e^{-\boldsymbol{\beta}^{\mathrm{T}} x}} \tag{3}$$

在换机预测应用中，假设给定 n 个因素 x=(x1,x2,⋯,xn)，设条件概率 P=P(y=1|x)为换机事件 y 相对于因素 x 发生的概率，用 Logistic 函数表示为

$$\begin{aligned} P=h_\beta(x)&=\frac{1}{1+e^{-\boldsymbol{\beta}^{\mathrm{T}} x}} \\ &=\frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}} \\ &=\frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n}} \end{aligned} \tag{4}$$

根据式(4)，可建立换机预测模型。其中，P 表示用户换机的概率，x1、x2、……、xn 表示换机的影响因素如手机品牌，上网天数，平均流量等。

2.1 逻辑回归模型工作流程

逻辑回归模型工作流程主要包括模型训练与模型预测两个阶段，具体流程如下。

1）模型训练流程

a)对选取的训练数据集进行数据预处理，包括对数据样本的去噪、剔重和筛选。

b)通过相关性分析得到对模型预测结果影响显著的变量，并将其作为模型最终的输入变量。

c)训练数据集利用逻辑回归的 fit()方法生成逻辑回归模型。

2)模型预测流程

a)针对测试数据集进行数据预处理，并选取对预测结果影响显著的变量输入模型，该模型为经过训练数据集所生成的逻辑回归模型。

b)调用训练后的逻辑回归模型的 transform()方法对测试数据进行分析，输出潜在换机用户。

c)输出并保存预测结果。

3　换机预测的实现与应用

换机预测系统是基于 Hadoop 平台实现的，其数据预处理过程包括数据字段提取，标准化处理等，主要是在 Hive 上完成，而换机预测模型的建立、预测等工作主要是在 Spark 组件上完成。

3.1　数据准备阶段

原始数据集 O 包括外部数据和信令数据，涵盖通信行为信息、网络行为信息、用户终端信息、用户属性信息、用户消费信息、boss、经分信息等多维度数据，共 146 字段。表 1 展示了原始数据集 O 的字段信息，由于字段较多，只选取部分展示。

表 1　原始数据集字段信息

用户终端信息	终端品牌	IMEI 在网月数
	终端网络类型	……
用户消费信息	当月消费	漫游通话费
	最近三个月平均消费	……
通信行为信息	通话常驻区县	通话常驻区域
	通话常驻片区（分局）	……
网络行为信息	上网套餐	视频流量
	上网叠加套餐	……
……	……	……

由表 1 可知，原始数据集 O 涵盖了不同维度的字段信息，也是换机预测模型的数据源。在系统实现过程中，需要建立 boss、信令等数据宽表，存储于 HDFS 中，作为模型输入的数据基础宽表。

3.2　数据预处理阶段

3.2.1 特征字段提取

原始数据集 O 所含字段较多，其中有些字段与换机预测关联较小，因此需要对宽表数据进行特征字段提取，留下可能对换机产生影响的变量字段。根据经验法，从 BOSS、信令等宽表中筛选出 17 个相关特征变量字段，包括手机号（usr_nbr）、手机品牌（phone_brand）、当月流量（gprs_flux）、三个月平均流量（flow_avg3）、4g 流量（flow_4g）等字段。

3.2.2 相关性分析

相关性分析即对提取字段进行两两间的相关性分析，本文借助了 spass statistics 工具，对所有字段做双变量相关性分析，采用 pearson 相关系数算法，所得相关性分析结果部分如表 2 所示。

从表 2 中可以看出，gprs_flux 与 flow_avg3 字段的相关系数为 0.819，gprs_flux 与 flow_4g 字段相关性系数为 0.963，flow_4g 与 flow_avg3 字段的相关系数为 0.863，相关性较大，当两字段相关系数大于 0.8 时，选取其中一个作为模型的输入，因此模型输入字段去掉 gprs_flux 与 flow_avg3 字段。

3.2.3 数据标准化处理

数据标准化处理先把非数值型数据进行转换，再对数据进行归一化处理。本文对于模型各输入字段处理方法不同，如对于市场细分（segment_type），将城区的转换为 1，其他的转换为 0；上网天数（gprs_days）为空的转换为 0；针对手机品牌（phone_brand）字段，取值范围为 0 至 20，即苹果、欧珀、步步高、华为、MIUI、三星、联想、金立、诺基亚等手机终端品牌从 20 至 1 依次递减排序，当字段内容不属于 top20 的品牌，或手机品牌字段为空时都转换为 0。

表 2　部分相关性分析结果

相关性分析	usr_nbr	phone_brand	gprs_flux	flow_avg3	flow_2g	flow_3g	flow_4g
usr_nbr	1	0.327	0.197	0.028	0.108	0.282	0.239
phone_brand	0.327	1	0.206	-0.018	0.253	0.095	0.086
gprs_flux	0.197	0.206	1	**0.819**	0.157	0.036	**0.963**
flow_avg3	0.028	-0.018	**0.819**	1	0.568	0.264	**0.863**
flow_2g	0.108	0.253	0.157	0.568	1	0.237	0.197
flow_3g	0.282	0.095	0.036	0.264	0.237	1	0.149
flow_4g	0.239	0.086	**0.963**	**0.863**	0.197	0.149	1
innet_months	0.182	0.156	0.065	0.238	0.234	0.438	0.451

数据归一化处理是为了消除指不同量纲间对数据分析结果的影响，其处理方法为：新数据=（原数据-均值）/标准差。以在网月数（innet_months）与市场细分（segment_type）为例，归一化处理的代码如下：

```
(nvl(a.innet_months,0)-b.avg_innet_months)
/b.stddev_innet_months innet_months,
(nvl(a.segment_type,0)-b.avg_segment_type)
/b.stddev_segment_type segment_type ,
```

3.3 数据建模实现

数据建模是在 Hadoop 平台 spark 组件上实现的，其主要建模实现步骤如下：

a) 加载训练数据 traindata，其路径为 “user/test/train201706.csv”，并设置分割符以及划分属性列和标识列。

b) 建立逻辑回归对象 val lr = new LogisticRegression()。

c) 重新设置逻辑回归对象参数 lr.setMaxIter(10).setRegParam(0.01)，最大迭代次数 10，正则化参数 0.01。

d)训练模型。根据设定的模型参数，调用逻辑回归的 fit()方法拟合训练得到模型 model2，可以通过模型 model2 预测用户是否换机。

e) 加载测试数据 testdata，其路径为 “user/test/unchanges201707.csv”，并设置分割符以及划分属性列和标识列。

f)调用预测模型 model2 的 transform()方法，对测试数据 testdata 进行换机预测。

g)测试数据 testdata 通过预测模型后得到预测结果，输出保存至 res.csv 文件中。

模型实现的关键代码如下：

```
val traindata = sc.textFile("/user/test/train201706.csv")
val training = sqlContext.createDataFrame(......).toDF("label", "features")
val lr = new LogisticRegression()
lr.setMaxIter(10).setRegParam(0.01)
......
val model2 = lr.fit(training, paramMapCombined)
val testdata=sc.textFile("/user/test/unchanges201707.csv")
......
val res=model2.transform(test).select("isdn","features", "myProbability", "prediction")
res.save(path="/user/test/res",source="json")
```

3.4 结果分析

在换机预测模型完成之后，本文选取了某西部城市 2017 年 7 月部分用户数据进行系统测试，测试对象为 200000 名未换机用户，将该数据导入换机预测系统后得到结果如表 3 所示，其中，“134****4763”是用户号码，“b”代表预测用户是潜在换机用户，“a”代表预测用户为非潜在换机用户。

表 3 换机预测模型结果

134****4763	b
134****6829	b
134****0019	a
……	…

通过换机预测模型后，得到预测结果，进一步，利用大数据平台统计潜在换机用户人数，关键代码如下：

```
---14002
select count(a.usr_nbr) from cyhcmc_temp.hj_jychange_test a,hj_change_201707 b
where a.usr_nbr = b.usr_nbr and a.flag = 'b';
---19720
select count(a.usr_nbr) from cyhcmc_temp.hj_jychange_test a
where a.flag = 'b';
```

从统计结果可以看出，待测试 200000 名用户中，预测所得潜在换机用户数为 19720，其中有 14002 用户是与实际相匹配的，实际上 7 月份换机用户数为 18216，预测准确率为 14002/19720=71%。系统测试中，从开始数据处理到最终模型完成预测共花费时间约 5.3s。现有换机预测研究准确率多为 68%-74%左右，本文实现的换机预测系统准确率与现有研究相当，但相比于传统仿真或理论研究的实现方式，该换机预测系统性能有较高的提升，可及时预测出潜在换机用户。该西部城市运营商针对该换机预测系统测预测的潜在用户，进行了精准营销，成功营销人数达 5248，成功营销率为 5248/19457=27.0%，相比于该西部城市运营商之前随机抽取用户 4%~5%的营销成功率有了更大的提升。

由运营商实际成功营销数据表明，该换机预测系统可以预测出潜在的换机用户，运营商仅仅对潜在换机用户实施精准营销，从人力投入方面降低了营销成本，并且最终提高手机终端的营销成功率。此外，该换机预测系统还可以掌握用户终端使用的行为习惯、换机前后的消费变化，结合用户消费行为数据，为后续其他业务（如套餐推荐）提供参考。

4 结束语

本文从理论与实际结合的角度研究换机预测，并应用 Hadoop 平台上的 Spark 组件实现了换机预测系统，能较好的识别出潜在的换机用户。但换机预测模型仍然存在一些问题，例如选取特征字段的方法是借助专家经验法，可能会忽略一些影响换机预测的其他字段；模型的预测结果与实际结果仍存在一些误差，可能是用户是否换机会受到一些偶然因素的影响。在接下来的研究中，会尝试用聚类算法选取相关特征字段，进一步提高换机预测模型的准确率。

参考文献：

[1] Ma J, Zhao S, Ma J, et al. Research on precision marketing data source system based on big data [J]. International Journal of Advanced Media & Communication, 2017, 7 (2): 93.

[2] You Z, Si Y W, Zhang D, et al. A decision-making framework for precision marketing [J]. Expert Systems with Applications, 2015, 42 (7): 3357-3367.

[3] 刘力凯, 王国胤, 邓维斌. 优势关系粗糙集的移动用户换机预测方法 [J]. 小型微型计算机系统, 2015, 36 (8): 1789-1794.

[4] 刘畅. 基于 Cox 回归模型的用户终端换机研究 [J]. 电子科学技术,

2016, 3 (4): 418-421.

[5] 熊冰妍，王国胤，邓维斌. 分级式代价敏感决策树及其在手机换机预测中的应用 [J]. 山东大学学报：工学版, 2015, 45 (5): 36-42.

[6] Yang D, Wu Z, Wang X, et al. Predicting replacement of smartphones with mobile app usage [C]// Proc of International Conference on Web Information Systems Engineering. [S. l.] : Springer International Publishing, 2016: 343-351.

[7] Ghazi M R, Gangodkar D. Hadoop, MapReduce and HDFS: a developers perspective [J]. Procedia Computer Science, 2015, 48: 45-50.

[8] Edition S. Applied logistic regression analysis [J]. Technometrics, 2017, 38 (2): 184-186.

[9] Wallenstein S, Hodge S E, Weston A. Logistic regression model for analyzing extended haplotype data [J]. Genetic Epidemiology, 2015, 15 (2): 173-181.

[10] Zhang Z. Model building strategy for logistic regression: purposeful selection. [J]. Annals of Translational Medicine, 2016, 4 (6): 111.

[11] Zhang S, Zhang L, Qiu K, et al. Variable selection in logistic regression model [J]. Chinese Journal of Electronics, 2015, 24 (4): 813-817.

[12] Olive D J. Linear regression analysis [J]. Technometrics, 2014, 45 (4): 362-363.

[13] Abuella M, Chowdhury B. Solar power probabilistic forecasting by using multiple linear regression analysis [C]// Proc of Southeastcon. 2015: 1-5.